

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2001-318948
(P2001-318948A)

(43) 公開日 平成13年11月16日 (2001. 11. 16)

(51) Int.Cl. ⁷	識別記号	F I	データベース (参考)
G 0 6 F 17/30	3 5 0	G 0 6 F 17/30	3 5 0 C 5 B 0 7 5
	1 7 0		1 7 0 A
	3 4 0		3 4 0 B
	3 8 0		3 8 0 Z

審査請求 未請求 請求項の数13 O L (全 18 頁)

(21) 出願番号 特願2000-142232(P2000-142232)

(22) 出願日 平成12年5月9日 (2000. 5. 9)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 菅谷 奈津子

神奈川県川崎市幸区鹿島田890番地 株式
会社日立製作所ビジネスソリューション開
発本部内

(72) 発明者 多田 勝己

神奈川県川崎市幸区鹿島田890番地 株式
会社日立製作所ビジネスソリューション開
発本部内

(74) 代理人 100075096

弁理士 作田 康夫

最終頁に続く

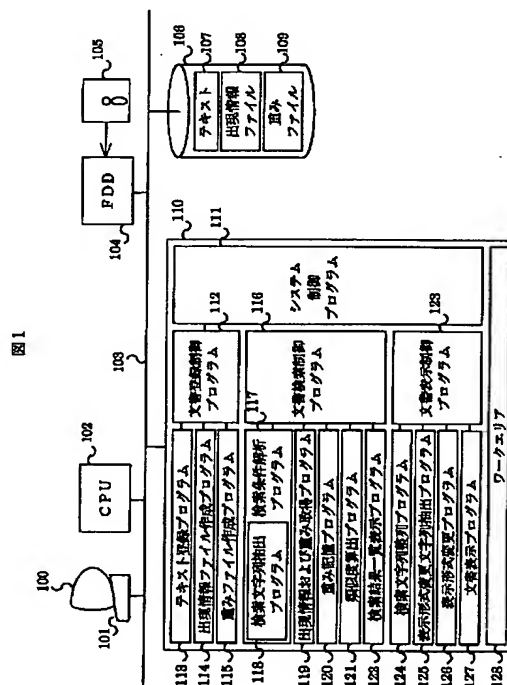
(54) 【発明の名称】 文書検索方法及び装置並びにその処理プログラムを記憶した媒体

(57) 【要約】

【課題】 文書や長い文章からなる条件で検索された文書を表示する場合でも、重要箇所の認識が容易な文書表示インターフェースを持つ文書検索システムを提供する。

【解決手段】 登録時には、登録対象テキストから抽出した所定の文字列と出現情報を出現情報ファイルに格納し、所定の方法で算出した文字列の重みを重みファイルに格納する。検索時には、指定された検索条件から所定の文字列を抽出し、出現情報ファイルと重みファイルから取得した文字列の出現情報と重みを用いて検索条件とデータベース内テキストとの類似度を算出する。文書表示時には、検索に用いた文字列から重みの高い文字列を抽出し、表示対象テキストの中で抽出した文字列が含まれる部分の表示形式を変更し、テキストを表示する。

【効果】 ユーザが表示文書の重要箇所を瞬時に認識でき、所望の文書か否かを高速に判別することが可能となる。



【特許請求の範囲】

【請求項1】文書情報を文字コードデータとして蓄積したテキストデータベースを対象として、検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する文書検索方法において、該類似度を算出する際に用いた情報の中で、重要な情報を選択し、該情報を表示するステップを有することを特徴とした文書検索方法。

【請求項2】文書情報を文字コードデータとして蓄積したテキストデータベースを対象として、登録対象テキストデータから所定の部分文字列と該部分文字列の該登録対象テキストデータにおける出現情報を抽出し、出現情報ファイルとして記憶する文書登録ステップと、検索条件から所定の部分文字列を抽出し、該出現情報ファイルから取得した該部分文字列の出現情報を用いて、検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する文書検索ステップを有する文書検索方法において、該部分文字列から重要な部分文字列を選択し、該類似度を算出する際に用いた情報の中で、該重要な部分文字列に関する情報を表示する重要情報表示ステップを有することを特徴とした文書検索方法。

【請求項3】請求項2記載の文書検索方法において、前記重要情報表示ステップは、前記部分文字列の重要度を所定の方法で算出し、該重要度の降順に所定の数の部分文字列を抽出する高重要度部分文字列抽出ステップと、該高重要度部分文字列抽出ステップで抽出した部分文字列に関する前記出現情報あるいは前記類似度の算出に寄与した度合いを表示する高重要度情報表示ステップを有し、前記重要な部分文字列とは、該高重要度部分文字列抽出ステップで、該重要度の降順に所定の数だけ抽出した部分文字列であり、前記重要な部分文字列に関する情報とは、該重要な部分文字列に関する該出現情報あるいは該類似度の算出に寄与した度合いを含むことを特徴とした文書検索方法。

【請求項4】請求項2記載の文書検索方法において、前記文書検索ステップは、検索条件から所定の部分文字列を抽出し、情報を表示する部分文字列を該部分文字列の中からユーザに選択させる情報表示部分文字列選択ステップと、前記出現情報ファイルから取得した該部分文字列の出現情報を用いて、検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する類似度算出ステップを有し、前記重要情報表示ステップは、該情報表示部分文字列選択ステップにおいて選択された部分文字列を取得する選択部分文字列取得ステップと、該選択部分文字列取得

ステップで取得した部分文字列に関する前記出現情報あるいは前記類似度の算出に寄与した度合いを表示する選択情報表示ステップを有し、

前記重要な部分文字列とは、該情報表示部分文字列選択ステップにおいて、選択された部分文字列であり、前記重要な部分文字列に関する情報とは、該重要な部分文字列に関する該出現情報あるいは該類似度の算出に寄与した度合いを含むことを特徴とした文書検索方法。

【請求項5】請求項2記載の文書検索方法において、前記文書検索ステップは、検索条件から所定の部分文字列を抽出し、該部分文字列に対して部分文字列の追加あるいは削除をユーザに実施させる部分文字列編集ステップと、

該部分文字列編集ステップで編集された該部分文字列の出現情報を、前記出現情報ファイルから取得し、該出現情報を用いて、検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する編集後類似度算出ステップを有し、前記重要情報表示ステップは、該部分文字列編集ステップにおいて追加された部分文字列を取得する追加部分文字列取得ステップと、該追加部分文字列取得ステップで取得した部分文字列に関する前記出現情報あるいは前記類似度の算出に寄与した度合いを表示する追加情報表示ステップを有し、

前記重要な部分文字列とは、該部分文字列編集ステップにおいて、追加された部分文字列であり、前記重要な部分文字列に関する情報とは、該重要な部分文字列に関する該出現情報あるいは該類似度の算出に寄与した度合いを含むことを特徴とした文書検索方法。

【請求項6】請求項2記載の文書検索方法において、前記重要情報表示ステップは、前記部分文字列から前記類似度の算出に寄与した順に所定の数の部分文字列を抽出する寄与部分文字列抽出ステップと、該寄与部分文字列抽出ステップで抽出した部分文字列に関する前記出現情報あるいは該類似度の算出に寄与した度合いを表示する寄与信息表示ステップを有し、前記重要な部分文字列とは、該寄与部分文字列抽出ステップで、該類似度の算出に寄与した順に所定の数だけ抽出した部分文字列であり、

前記重要な部分文字列に関する情報とは、該重要な部分文字列に関する該出現情報あるいは該類似度の算出に寄与した度合いを含むことを特徴とした文書検索方法。

【請求項7】文書情報を文字コードデータとして蓄積したテキストデータベースを対象として、登録対象テキストデータから所定の部分文字列と該部分文字列の該登録対象テキストデータにおける出現情報を抽出し、出現情報ファイルとして記憶する文書登録ステップと、検索条件から所定の部分文字列を抽出し、該出現情報ファイルから取得した該部分文字列の出現情報を用いて、検索条件の内容とテキストデータベース中の

10

20

30

40

50

テキストの内容が類似する度合いを類似度として所定の方法で算出する文書検索ステップと、
 該文書検索ステップで該類似度を算出したテキストの中でユーザに指定されたテキストを表示する文書表示ステップを有する文書検索方法において、
 該文書表示ステップは、該部分文字列から重要な部分文字列を選択し、該指定されたテキストの中で、該重要な部分文字列が含まれる部分の表示形式を変更する表示形式変更ステップを有することを特徴とした文書検索方法。

【請求項8】請求項7記載の文書検索方法において、
 前記文書登録ステップは、前記登録対象テキストデータから所定の部分文字列と該部分文字列の該登録対象テキストデータにおける出現情報を抽出し、出現情報ファイルとして記憶する出現情報ファイル作成ステップと、
 該部分文字列の重要度を所定の方法で算出し、重要度ファイルとして記憶する重要度ファイル作成ステップを有し、
 前記文書表示ステップは、該部分文字列の重要度を該重要度ファイルから取得し、該取得した重要度の降順に所定の数の部分文字列を抽出する重要部分文字列抽出ステップと、
 前記指定されたテキストの中で、該重要部分文字列抽出ステップで抽出した部分文字列が含まれる部分の表示形式を変更する重要部分文字列表示形式変更ステップを有し、
 前記重要な部分文字列とは、該重要部分文字列抽出ステップで、該重要度の降順に所定の数だけ抽出した部分文字列であることを特徴とした文書検索方法。

【請求項9】請求項7記載の文書検索方法において、
 前記文書検索ステップは、検索条件から所定の部分文字列を抽出し、表示形式を変更する部分文字列を該部分文字列の中からユーザに選択させる表示形式変更部分文字列選択ステップと、
 前記出現情報ファイルから取得した該部分文字列の出現情報を用いて、検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する類似度算出ステップを有し、
 前記文書表示ステップは、該表示形式変更部分文字列選択ステップにおいて選択された部分文字列を取得する選択部分文字列取得ステップと、
 前記指定されたテキストの中で、該選択部分文字列取得ステップで取得した部分文字列が含まれる部分の表示形式を変更する選択部分文字列表示形式変更ステップを有し、
 前記重要な部分文字列とは、該表示形式変更部分文字列選択ステップにおいて選択された該表示形式を変更する部分文字列であることを特徴とした文書検索方法。

【請求項10】請求項7記載の文書検索方法において、
 前記文書検索ステップは、検索条件から所定の部分文

字列を抽出し、該部分文字列に対しユーザに部分文字列の追加あるいは削除を実行させる部分文字列編集ステップと、

該部分文字列編集ステップで編集された該部分文字列の出現情報を、前記出現情報ファイルから取得し、該出現情報を用いて、検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する編集後類似度算出ステップを有し、
 前記文書表示ステップは、該部分文字列編集ステップにおいて追加された部分文字列を取得する追加部分文字列取得ステップと、

前記指定されたテキストの中で、該追加部分文字列取得ステップで取得した部分文字列が含まれる部分の表示形式を変更する追加部分文字列表示形式変更ステップを有し、

前記重要な部分文字列とは、該部分文字列編集ステップにおいて追加された該部分文字列であることを特徴とした文書検索方法。

【請求項11】請求項7記載の文書検索方法において、
 前記文書表示ステップは、前記部分文字列から前記類似度の算出に寄与した順に所定の数の部分文字列を抽出する寄与部分文字列抽出ステップと、
 前記指定されたテキストの中で、該寄与部分文字列抽出ステップで抽出した部分文字列が含まれる部分の表示形式を変更する寄与部分文字列表示形式変更ステップを有し、
 前記重要な部分文字列とは、該寄与部分文字列抽出ステップで、該類似度の算出に寄与した順に所定の数だけ抽出した部分文字列であることを特徴とした文書検索方法。

【請求項12】文書情報を文字コードデータとして蓄積したテキストデータベースを対象として、
 登録対象テキストデータから所定の部分文字列と該部分文字列の該登録対象テキストデータにおける出現情報を抽出し、出現情報ファイルとして記憶する文書登録手段と、

検索条件から所定の部分文字列を抽出し、該出現情報ファイルから取得した該部分文字列の出現情報を用いて、検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する文書検索手段と、

該文書検索手段で該類似度を算出したテキストの中でユーザに指定されたテキストを表示する文書表示手段を備える文書検索装置において、

該文書表示手段は、該部分文字列から重要な部分文字列を選択し、該指定されたテキストの中で該重要な部分文字列が含まれる部分の表示形式を変更する表示形式変更手段を備えることを特徴とした文書検索装置。

【請求項13】文書情報を文字コードデータとして蓄積したテキストデータベースを対象として、
 登録対象テキストデータから所定の部分文字列と該部分

10

20

30

40

50

文字列の該登録対象テキストデータにおける出現情報を抽出し、出現情報ファイルとして記憶する文書登録モジュールと、

検索条件から所定の部分文字列を抽出し、該出現情報ファイルから取得した該部分文字列の出現情報を用いて、検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する文書検索モジュールと、

該文書検索モジュールで該類似度を算出したテキストの中でユーザに指定されたテキストを表示する文書表示モジュールを含む文書検索システムを構築するためのプログラムを格納した記憶媒体において、

前記プログラムにおいて、該文書表示モジュールは、該部分文字列から重要な部分文字列を選択し、該指定されたテキストの中で該重要な部分文字列が含まれる部分の表示形式を変更する表示形式変更モジュールを含むことを特徴とした記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、入力した条件に基づき登録文書を検索し、その検索した文書に関する情報を表示する文書検索技術に係る。

【0002】

【従来の技術】近年、ワードプロセッサ等により作成される電子化文書が増大しており、今後も増大していくことが見込まれる。このため、文書検索の対象となるデータベースも大規模になってきている。その結果、文書検索によって得られる検索結果としての文書集合も大型化し、この中からユーザが本当に欲しい文書を探し出すことが困難になってきている。

【0003】この問題を改善するため、従来よりランキング技術が提案されている。

【0004】ランキング技術については、「Ranking Algorithms」(Donna Harman著、Information Retrieval, p.363-392)に具体的に記載されている。以下、これを従来技術1と呼ぶ。従来技術1は、ユーザが指定した検索条件(文章、文書、または複数の単語の列)の内容と類似する可能性を示す指標を算出する技術である。以下、その内容の一例を、図2を用いて説明する。

【0005】検索は、簡単なベクトル演算によって実現される。このベクトルの要素は、データベース中に出現する全ての単語を重複排除したもの(但し、ストップワード等は除く)である。図2の例では、(factors, information, help, human, operation, retrieval, systems)が要素となっている。検索条件にその要素が存在すれば「1」を、存在しなければ「0」を該当位置に立てる。この結果、検索条件のベクトルQ0が作成される。すなわち、「human factors in information retrieval systems」という検索条件に対し、(1, 1, 0, 1, 0, 1, 1)というベクトルQ0が作成される。

【0006】データベース中の文書に対しても同様に文書のベクトルが作成される。「factors」「information」「human」「retrieval」が含まれる文書1に対し、ベクトルV1(1, 1, 0, 1, 0, 1, 0)が作成される。また、「factors」「help」「human」「systems」が含まれる文書2に対し、ベクトルV2(1, 0, 1, 1, 0, 0, 1)が作成される。さらに、「factors」「operation」「systems」が含まれる文書3に対し、ベクトルV3(1, 0, 0, 0, 1, 0, 1)が作成される。

【0007】ここで、ランキングに用いられる得点は、検索条件のベクトルQ0と文書のベクトルVi(i=1, 2, 3)とのベクトル積Vi・Q0をとることによって算出される。その結果、文書1が4点、文書2が3点、文書3が2点となる。この得点はシステムが判断した検索条件との類似度であり、この得点が高い文書ほど検索条件の内容と類似する可能性が高いことを示している。

【0008】なお、ベクトルの要素を「1」「0」ではなく、単語の重み(単語の出現頻度や文書データベース中の出現の偏り具合等で算出)等で表現することも可能である。例えば、「factors」の重み2、「information」の重み3、「human」の重み5、「retrieval」の重み3を用いて、文書1に対し、ベクトルV'1(2, 3, 0, 5, 0, 3, 0)が作成される。また同様に、「factors」の重み2、「help」の重み4、「human」の重み5、「systems」の重み1を用いて、文書2に対し、ベクトルV'2(2, 0, 4, 5, 0, 0, 1)が作成される。さらに、「factors」の重み2、「operation」の重み2、「systems」の重み1を用いて、文書3に対し、ベクトルV'3(2, 0, 0, 0, 2, 0, 1)が作成される。

【0009】これらのベクトルV'iと検索条件式のベクトルQ0とのベクトル積V'i・Q0をとることによって、各文書の得点が算出される。その結果、文書1が13点、文書2が8点、文書3が3点となる。この得点は単語の重み、すなわち単語の重要度を考慮してシステムが判断した検索条件との類似度であり、この得点が高い文書ほど検索条件の内容と類似する可能性が高いことを示している。つまり、文書1が最も検索条件の内容と類似する可能性が高いとの結果を得る。

【0010】この従来技術1では、検索条件の内容と類似する可能性を示す指標を算出しており、この指標に従って文書を閲覧することにより、大規模な文書データベースの中から所望の文書を高速に探し出すことできる可能性がある。しかし、得られた検索結果文書が本当に欲しい文書であるか否かは、実際に文書の内容をユーザが読んで判別しなければならない。このように、検索結果として得られた文書が所望の文書であるか否かを瞬時に判別することを支援するための技術として、文書のハイ

ライト表示技術がある。以下、これを従来技術2と呼ぶ。

【0011】従来技術2は、検索結果として得られた文書の内容を表示する際に、ユーザが指定した検索条件中の文字列が含まれる部分を、他の文字列部分と異なる表示形式で表示する（以下、ハイライトと呼ぶ）というものである。ここでいう表示形式とは、例えば色、サイズ、フォントやスタイル（太字や斜体）等である。検索条件中の文字列が含まれる部分を他の文字列部分と異なる表示形式にすることにより、その単語が含まれる位置を瞬時に認識することができ、文書を先頭から読む場合よりも高速に所望の文書であるか否かを判別することが可能となる。

【0012】

【発明が解決しようとする課題】従来技術1に示すランキング技術で用いられるベクトルの要素としては、単語が用いられることが多い。英語のように単語が分かち書きされている言語の場合には、ストップワード（in, theなど）以外の全ての単語が用いられる。日本語のように分かち書きされていない言語の場合には、文字種の異なるところで分割して得られる文字列や、n文字（nは予め定められた1以上の整数）の連続する文字列、または辞書等を参照して抽出した単語等が用いられる。その結果、文書や長い文章を指定して検索を行い、その結果として得られた文書を従来技術2に示すハイライト表示技術で表示する際には、ハイライト対象となる文字列が多くなり、かえって重要な部分が見つけれなくなるとい

う問題がある。

【0013】以下、図3を用いて新聞記事データベースを例として説明する。本例では、検索条件としてサッカーのワールドカップの会場誘致に関する新聞記事文書を指定して検索を行っている。

【0014】まず、検索条件として指定された文書「サッカーW杯試合会場、来月決定、選定の権限は協会に。日韓共催の二〇〇二年サッカーワールドカップ開催準備委員会は二十九日、開催地の候補である十五自治体の最高責任者らを集めて知事・市長会議を開いた。国際サッカー連盟（FIFA）が国内会場数を…」から検索に用いる文字列を抽出する。図3に示す例では、辞書等を参照して抽出した名詞、片仮名およびアルファベットを検索に用いる文字列として抽出している。この結果、検索条件から「サッカー、W杯、試合、会場、来月、決定、選定、権限、協会、日、韓、共催、ワールドカップ、開催、準備、委員会、地、候補、…」が抽出され、これらの文字列を用いて検索が行われる。その結果、検索条件の内容と類似する可能性を示す指標が算出され文書の一覧とともに出力される。そこで、検索条件の内容と類似する可能性が高い、すなわち最も得点の高い文書から閲覧し、本当に欲しい文書であるか否かをユーザが確認することになる。ここで、従来技術2のようなハイライト

表示技術を利用すると、検索条件中の文字列が含まれる位置を瞬時に認識できるので、文書を先頭から読む場合よりも高速に所望の文書であるか否かを判別することができる。しかし、図3に示すように、文書や長い文章を指定して検索を行い、その結果として得られた文書を表示する際には、検索に用いる文字列が多いためにハイライト箇所（図3の例では、サイズを大きくし、斜体、強調を施している）が多くなり、かえって重要な部分が見つけれなくなってしまう。

【0015】こうした問題に対し、本発明では、ユーザが所望とする文書か否かを容易に判別することが可能となる文書情報表示機能を実現することを目的とする。

【0016】

【課題を解決するための手段】上記課題を改善するために、本発明では、以下の処理ステップを有する。

【0017】すなわち、文書情報を文字コードデータとして蓄積したテキストデータベースを対象として、ユーザが指定した検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する文書検索ステップと、文書検索ステップにおいて類似度を算出する際に用いた情報の中で、重要な情報を選択し、該情報を表示する文書表示ステップを有する。

【0018】上記文書検索方法を用いた本発明の原理を、以下に説明する。文書の検索時にユーザから検索条件として文章や文書が指定された場合には、上述した文書検索ステップを実行し、ユーザが指定した検索条件の内容とテキストデータベース中のテキストの内容が類似する度合いを類似度として所定の方法で算出する。以下、文書検索ステップの処理内容の例を示す。まず、指定された検索条件（以下、検索条件文書と呼ぶ）から、予め定められた文字列を抽出する。この文字列としては、英語等の分かち書きされている言語であれば単語、そうでなければ文字種の異なるところで分割して得られる文字列や、n文字（nは予め定められた1以上の整数）の連続する文字列、または辞書等を参照して抽出した単語等を用いる。検索条件文書として「サッカーW杯試合会場、来月決定、選定の権限は協会に。日韓共催の二〇〇二年サッカーワールドカップ開催準備委員会は二十九日、開催地の候補である十五自治体の最高責任者らを集めて知事・市長会議を開いた。国際サッカー連盟（FIFA）が国内会場数を…」が指定され、文字列として辞書等を参照して抽出した名詞、片仮名およびアルファベットを用いる場合には、図3に示すように、「サッカー、W杯、試合、会場、来月、決定、選定、権限、協会、日、韓、共催、ワールドカップ、開催、準備、委員会、地、候補、…」が抽出される。そして、これらの文字列のテキストデータベースにおける出現情報を抽出する。この出現情報としては、使用する検索方式により異なるが、文字列の出現する文書の番号や出現位置、出

現回数等を用いる。従来技術1の場合には、文書のベクトルを作成する際に必要となる文字列の出現文書番号や出現回数を用いる。さらに、この出現情報から所定の方法を用いて文字列の重みを算出する。この重みは、使用する検索方式によりその算出方法が異なるが、文字列の出現頻度や文書データベース中の出現の偏り具合等を用いて算出する。出現の偏り具合から算出される重みとして一般的に用いられているのは、「Ranking Algorithms」(Donna Harman著、Information Retrieval、p.363-392)にも記載されているIDF(Inverted Document Frequency)である。IDFは多くの文書に出現する文字列はストップワードである可能性が高く、重要度は低いという考え方から生み出された重みである。類似度はこの出現情報および重みを用いて、予め定められた算出方法で算出する。この算出方法は使用する検索方式により異なるが、例えば、図2に示す従来技術1で用いられている簡単なベクトル演算によっても算出できる。そして、算出した類似度を検索結果一覧として表示する。

【0019】検索結果一覧上で選択した文書の表示要求がなされた場合には、上述した文書表示ステップを実行し、文書検索ステップにおいて類似度を算出する際に用いた情報の中で、重要な情報を選択し、該情報を表示する。以下、文書検索ステップの処理内容の例を示す。まず、文書検索ステップにおいて検索条件から抽出された文字列とその重みを、重みの降順に整列する。そして、整列した文字列の上位m個(mは予め定められた1以上の整数)を抽出する。このmの値はシステムで自動的に適切な値を設定しても良いし、ユーザに予め設定させておいても良い。また、文書表示毎に対話的にユーザに設定させ、適切な値を調整してもかまわない。図4に抽出した文字列の例を示す。本図に示す例では、mとして4を用いている。重みの降順に整列した文字列の上位4個を抽出することにより、「W杯」「サッカー」「ワールドカップ」「FIFA」が抽出される。次に、ユーザが表示対象として指定したテキスト(以下、表示文書と呼ぶ)の中で、抽出した文字列が含まれる部分の表示形式を変更し、表示文書を表示する。図4に示すように、抽出した「W杯」「サッカー」「ワールドカップ」「FIFA」が含まれる部分の表示形式(図4の例では、サイズを大きくし、斜体、強調を施している)を変更し、文書を表示する。こうすることにより、ユーザは文書中の重要箇所を一目で認識することができるようになる。本処理例において、上述した重要な情報とは重要な文字列に関するハイライト情報のことであり、重要な文字列とは重みの降順の上位m個(mは予め定められた1以上の整数)の文字列のことである。

【0020】以上説明したように、本方式では、検索条件の内容と類似する可能性を示す指標に対して影響する文字列、例えば重みの高いものから順に予め決められた数の文字列を選定し、その文字列に関する情報として、

例えばその文字列が含まれる部分だけ表示形式を変更して文書を表示する。こうすることにより、検索に用いられた文字列の中で重要な文字列に関する情報だけを表示することになるので、ユーザは文書中の重要箇所を瞬時に認識し、所望の文書か否かを高速に判別することができる。その結果、検索結果文書を閲覧する際のユーザインターフェースを向上することが可能となる。

【0021】

【発明の実施の形態】以下、本発明の第一の実施例について図1を用いて説明する。

【0022】本発明を適用した文書検索システムは、ディスプレイ100、キーボード101、中央演算処理装置(CPU)102、磁気ディスク装置106、フロッピディスクドライバ(FDD)104、主記憶装置110およびこれらを結ぶバス103から構成される。磁気ディスク装置106は二次記憶装置の一つであり、テキスト107、出現情報ファイル108および重みファイル109が格納される。フロッピディスク105に格納されている情報は、FDD104によりアクセスされる。

【0023】主記憶装置110には、システム制御プログラム111、文書登録制御プログラム112、テキスト登録プログラム113、出現情報ファイル作成プログラム114、重みファイル作成プログラム115、文書検索制御プログラム116、検索条件解析プログラム117、出現情報および重み取得プログラム119、重み記憶プログラム120、類似度算出プログラム121、検索結果一覧表示プログラム122、文書表示制御プログラム123、検索文字列整列プログラム124、表示形式変更文字列抽出プログラム125、表示形式変更プログラム126および文書表示プログラム127が格納されるとともにワークエリア128が確保される。検索文字列抽出プログラム118は検索条件解析プログラム117に含まれる。以上のプログラムは磁気ディスク装置106、フロッピディスク105などのコンピュータで読み書きできる記憶媒体に格納することもできる。

【0024】システム制御プログラム111はキーボード101からの指示を受け起動する。文書登録制御プログラム112はキーボード101からの文書登録指示により、システム制御プログラム111によって起動され、テキスト登録プログラム113、出現情報ファイル作成プログラム114および重みファイル作成プログラム115の制御を行う。文書検索制御プログラム116はキーボード101からの文書検索指示により、システム制御プログラム111によって起動され、検索条件解析プログラム117、出現情報および重み取得プログラム119、重み記憶プログラム120、類似度算出プログラム121および検索結果一覧表示プログラム122の制御を行う。文書表示制御プログラム123はキーボード101からの文書表示指示により、システム制御プ

プログラム111によって起動され、検索文字列整列プログラム124、表示形式変更文字列抽出プログラム125、表示形式変更プログラム126および文書表示プログラム127の制御を行う。

【0025】以下、本実施例における処理内容の概要を説明する。

【0026】文書を登録する際には、キーボード101からの文書登録指示により、システム制御プログラム111が文書登録制御プログラム112を起動し、文書登録制御プログラム112がテキスト登録プログラム113、出現情報ファイル作成プログラム114および重みファイル作成プログラム115による一連の文書登録処理を制御する。文書登録処理、すなわち文書登録制御プログラム112の処理内容を図5のPAD (Problem Analysis Diagram) 図に示す。文書登録制御プログラム112は、図5に示すように、まずステップ600で、テキスト登録プログラム113を起動し、FDD104に挿入されたフロッピディスク105から登録する文書のテキストデータを読み込み、これをテキスト107として磁気ディスク装置106に格納する。テキストデータはフロッピディスク105を用いて入力するだけに限らず、通信回線やCD-ROM装置(図1には示していない)等を用いて他の装置から入力するような構成をとることも可能である。次にステップ601で、出現情報ファイル作成プログラム114を起動して、ステップ600で読み込んだテキストデータから予め定められた文字列とその出現情報を抽出し、出現情報ファイル108として磁気ディスク装置106に格納する。ここで抽出する文字列としては、英語等の分かち書きされている言語であれば単語、そうでなければ文字種の異なるところで分割して得られる文字列や、n文字(nは予め定められた1以上の整数)の連続する文字列、または辞書等を参照して抽出した単語等を用いる。また出現情報としては、使用する検索方式により異なるが、文字列の出現する文書の番号や出現位置、出現回数等、検索に必要な情報を用いる。従来技術1の場合には、文書のベクトルを作成する際に必要となる文字列の出現文書番号や出現回数を用いる。最後にステップ602で、重みファイル作成プログラム115を起動し、ステップ601で抽出した文字列の重みを、予め定められた算出方法を用いて算出し、重みファイル109として磁気ディスク装置106に格納する。この重みは、使用する検索方式によりその算出方法が異なるが、文字列の出現頻度や文書データベース中の出現の偏り具合(一般的には「Ranking Algorithms」(Donna Harman著、Information Retrieval、p.363-392)記載のIDFを用いる)等を用いて算出する。以上で文書登録処理は終了する。

【0027】文書を検索する際には、キーボード101からの文書検索指示により、システム制御プログラム111が文書検索制御プログラム116を起動し、文書検

索制御プログラム116が検索条件解析プログラム117、出現情報および重み取得プログラム119、重み記憶プログラム120、類似度算出プログラム121および検索結果一覧表示プログラム122による一連の文書検索処理を制御する。文書検索制御プログラム116は、まず文書検索の前準備として、磁気ディスク装置106に格納されている出現情報ファイル108と重みファイル109を主メモリ110に確保されたワークエリア128に読み込む。そして、文書検索処理では、図6のPAD図に示すように、まずステップ700で、検索条件解析プログラム117を起動し、キーボード101から入力された検索条件を解析する。ここで、入力された検索条件には文章または文書が指定されていると、ステップ701で判断された場合には、ステップ702で検索文字列抽出プログラム118を用いて、検索条件として指定された文章または文書から、予め定められた文字列を抽出する。この文字列としては、文書登録制御プログラム112による文書登録処理での文字列抽出と同様に、英語等の分かち書きされている言語であれば単語、そうでなければ文字種の異なるところで分割して得られる文字列や、n文字(nは予め定められた1以上の整数)の連続する文字列、または辞書等を参照して抽出した単語等を用いる。また、入力された検索条件には複数の単語の列が指定されていると、ステップ701で判断された場合には、ステップ703で検索条件から単語の列を抽出する。次にステップ704で、出現情報および重み取得プログラム119を起動し、検索条件から抽出した文字列(あるいは単語)の出現情報および重みを、ワークエリア128に読み込んだ出現情報ファイル108と重みファイル109から取得する。この文字列と重みは、ステップ705で重み記憶プログラム120を起動し、ワークエリア128に記憶しておく。次にステップ706で、類似度算出プログラム121を起動し、ステップ704で取得した文字列の出現情報と重みを用いて、予め定められた算出方法で検索条件とテキストデータ間の類似度を算出する。この算出方法は使用する検索方式により異なるが、例えば、従来技術1で用いられている簡単なベクトル演算によっても算出できる。最後にステップ707で、検索結果一覧表示プログラム122を起動し、ステップ706で算出した類似度の降順にテキストデータを整列し、検索結果一覧として表示する。以上で文書検索処理は終了する。

【0028】文書を表示する際には、キーボード101からの文書表示指示により、システム制御プログラム111が文書表示制御プログラム123を起動し、文書表示制御プログラム123が検索文字列整列プログラム124、表示形式変更文字列抽出プログラム125、表示形式変更プログラム126および文書表示プログラム127による一連の文書表示処理を制御する。文書表示処理の処理内容を図7のPAD図に示す。文書表示制御プ

プログラム123は、図7に示すように、まずステップ800で、検索文字列整列プログラム124を起動し、文書検索制御プログラム116による文書検索処理でワークエリア128に記憶した文字列とその重みを、重みの降順に整列する。次にステップ801で、表示形式変更文字列抽出プログラム125を起動し、ステップ800で整列した文字列の上位m個（mは予め定められた1以上の整数）を抽出する。このmの値はシステムで自動的に適切な値を設定しても良いし、ユーザに予め設定させておいても良い。また、文書表示毎に対話的にユーザに設定させ、適切な値を調整してもかまわない。次にステップ802で、表示形式変更プログラム126を起動し、表示対象として指定された文書（以下、表示文書と呼ぶ）の中で、ステップ801で抽出した文字列が含まれる部分の表示形式を変更する。この表示形式の変更方法は従来技術2と同様である。最後にステップ803で、文書表示プログラム127を起動し、ステップ802で表示形式を変更した表示文書を表示して文書表示処理を終了する。

【0029】以上が本実施例における処理内容の概要である。

【0030】以下、本実施例の処理内容を具体例を用いて詳細に説明する。

【0031】文書登録制御プログラム112による文書登録処理の内容は図5に示す通りである。以下、具体的に説明する。まずステップ600で、テキスト登録プログラム113を起動し、FDD104に挿入されたフロッピーディスク105から登録する文書のテキストデータを読み込み、これをテキスト107として磁気ディスク装置106に格納する。次にステップ601で、出現情報ファイル作成プログラム114を起動して、ステップ600で読み込んだテキストデータから予め定められた文字列を抽出する。ここで抽出する文字列としては、英語等の分かち書きされている言語であれば単語、そうでなければ文字種の異なるところで分割して得られる文字列や、n文字（nは予め定められた1以上の整数）の連続する文字列、または辞書等を参照して抽出した単語等を用いる。さらに、抽出した文字列の出現情報を抽出し、出現情報ファイル108として磁気ディスク装置106に格納する。この出現情報としては、使用する検索方式により異なるが、文字列の出現する文書の番号や出現位置、出現回数等、検索に必要な情報を用いる。従来技術1の場合には、文書のベクトルを作成する際に必要となる文字列の出現文書番号や出現回数を用いる。図8に出現情報ファイル108の作成例を示す。本図に示す出現情報ファイル108には、辞書等を参照して抽出した単語が出現する文書の番号ならびにその文書における出現回数が格納されている。例えば、“W杯”という文字列は文書番号“1”の文書に“2”回出現しているので、その文書番号“1”ならびに出現回数“2”を出現

情報ファイルに格納する。最後にステップ602で、重みファイル作成プログラム115を起動し、ステップ601で抽出した文字列の重みを、予め定められた算出方法を用いて算出し、重みファイル109として磁気ディスク装置106に格納する。この重みは、使用する検索方式によりその算出方法が異なるが、文字列の出現頻度や文書データベース中の出現の偏り具合等を用いて算出する。出現の偏り具合から算出される重みとして一般的に用いられているのは、「Ranking Algorithms」（Donna Harman著、Information Retrieval、p.363-392）にも記載されているIDF（Inverse Document Frequency）である。IDFは多くの文書に出現する文字列はストップワードである可能性が高く、重要度は低いという考えから生み出された重みである。図8に示す重みファイル109には出現偏り具合を用いて算出した重みが格納されており、特定の文書にしか出現しない文字列“サッカー”の重みは高く、多くの文書に出現するストップワードに近い文字列“来月”の重みは小さくなる。以上で文書登録処理は終了する。

【0032】文書検索制御プログラム116による文書検索処理の内容は図6に示す通りである。以下、具体的に説明する。まずステップ700で、検索条件解析プログラム117を起動し、キーボード101から入力された検索条件を解析する。ここで、入力された検索条件には文章または文書が指定されていると、ステップ701で判断された場合には、ステップ702で検索文字列抽出プログラム118を用いて、検索条件として指定された文章または文書から、予め定められた文字列を抽出する。この文字列としては、文書登録制御プログラム112による文書登録処理での文字列抽出と同様に、英語等の分かち書きされている言語であれば単語、そうでなければ文字種の異なるところで分割して得られる文字列や、n文字（nは予め定められた1以上の整数）の連続する文字列、または辞書等を参照して抽出した単語等を用いる。例えば、検索条件として「サッカーW杯試合会場、来月決定、選定の権限は協会に。日韓共催の二〇〇二年サッカーワールドカップ開催準備委員会は二十九日、開催地の候補である十五自治体の最高責任者らを集めて知事・市長会議を開いた。国際サッカー連盟（FIFA）が国内会場数を…」が指定された場合の例を図3に示す。図3に示す例では、辞書等を参照して抽出した名詞、片仮名およびアルファベットを検索に用いる文字列として抽出している。この結果、検索条件から「サッカー、W杯、試合、会場、来月、決定、選定、権限、協会、日、韓、共催、ワールドカップ、開催、準備、委員会、地、候補、…」が抽出される。また、入力された検索条件には複数の単語の列が指定されていると、ステップ701で判断された場合には、ステップ703で検索条件から単語の列を抽出する。ここで、検索条件として、例えば、「“サッカー” [AND] “ワールドカッ

ブ」のように、「“サッカー”と“ワールドカップ”の両方の文字列が現れる文書を探せ」というような複数の単語の論理条件が指定された場合には、単語の列として“サッカー”と“ワールドカップ”を抽出する。次にステップ704で、出現情報および重み取得プログラム119を起動し、検索条件から抽出した文字列（あるいは単語）の出現情報および重みを、ワークエリア128に読み込んだ出現情報ファイル108と重みファイル109から取得する。例えば“サッカー”という文字列の出現情報と重みを、図8に示す出現情報ファイル108と重みファイル109から取得すると、文書番号“1”に“3”回出現するという出現情報と、“80”という重みが得られる。この文字列と重み、例えば“サッカー”の重み“80”は、ステップ705で重み記憶プログラム120を起動し、ワークエリア128に記憶しておく。次にステップ706で、類似度算出プログラム121を起動し、ステップ704で取得した文字列の出現情報と重みを用いて、予め定められた算出方法で検索条件とテキストデータ間の類似度を算出する。この算出方法は使用する検索方式により異なるが、例えば、従来技術1で用いられている簡単なベクトル演算によっても算出できる。最後にステップ707で、検索結果一覧表示プログラム122を起動し、ステップ706で算出した類似度の降順にテキストデータを整列し、図3に示すような検索結果一覧として表示する。この結果、文書番号“123”の文書が、指定された検索条件の内容と最も類似する可能性の高い文書であることが分かる。以上で文書検索処理は終了する。

【0033】文書表示制御プログラム123による文書表示処理の内容は図7に示す通りである。以下、具体的に説明する。まずステップ800で、検索文字列整列プログラム124を起動し、文書検索制御プログラム116による文書検索処理でワークエリア128に記憶した文字列とその重みを、重みの降順に整列する。この整列処理の処理内容を図4に示す。整列処理を行うことにより、文字列が重みの降順、すなわち、重み“80”の“W杯”、“サッカー”、重み“70”の“ワールドカップ”、重み“60”の“FIFA”、重み“50”の“会場”、…の順に並べられる。次にステップ801で、表示形式変更文字列抽出プログラム125を起動し、ステップ800で整列した文字列の上位m個（mは予め定められた1以上の整数）を抽出する。このmの値はシステムで自動的に適切な値を設定しても良いし、ユーザに予め設定させておいても良い。また、文書表示毎に対話的にユーザに設定させ、適切な値を調整してもかまわない。図4に示す例では、上位4個、すなわちmの値を4としている。その結果、表示形式を変更する文字列として“W杯”、“サッカー”、“ワールドカップ”および“FIFA”が抽出される。そしてステップ802で、表示形式変更プログラム126を起動し、表示対

象として指定された文書（以下、表示文書と呼ぶ）の中で、ステップ801で抽出した文字列が含まれる部分の表示形式を変更する。この表示形式の変更方法は従来技術2と同様である。最後にステップ803で、文書表示プログラム127を起動し、ステップ802で表示形式を変更した表示文書を表示する。図4に文書番号“123”、文書番号“003”の文書表示例を示す。本図に示す例では、表示形式を変更する文字列“W杯”、“サッカー”、“ワールドカップ”および“FIFA”が含まれる部分のサイズが大きく、斜体、強調が施されて表示されている。以上で、文書表示処理を終了する。

【0034】以上説明したように、本実施例では、検索に用いられた文字列の中から、重要な文字列を選定し、その文字列が含まれる部分の表示形式のみを変更している。文書や長い文章を指定して検索を行い、その結果として得られた文書を表示する際に、検索に用いた文字列全ての表示形式を変更すると、図3に示すように変更箇所が多くなり、かえって重要な部分が見つけにくくなるという問題があるが、本実施例に示す方法を用いることにより、重要度の高い箇所限定し、表示形式を変更することができる。

【0035】このことにより、ユーザは文書中の重要箇所を瞬時に認識し、所望の文書か否かを高速に判別することが可能となる。その結果、検索結果文書を閲覧する際のユーザインターフェースを向上することができる。

【0036】本実施例では、重みの高い文字列の表示形式を変更する場合の例について説明したが、表示形式を変えるだけでなく、重みの高い文字列に関する情報、例えば出現頻度等を一覧として表示することも上記と同様の処理で実現することが可能である。

【0037】また本実施例では、文字列を同一の表示形式に変更する場合の例について説明したが、文字列の重みの値に応じて、表示形式を変えることも上記と同様の処理で実現可能である。例えば、最も重みの高い文字列に対してはサイズを大きく、目立つ色に変更する。次に重みの高い文字列に対してはサイズのみを大きくすることにより、どの文字列がどの程度重要であるかを瞬時に認識することができる。

【0038】以下、本発明の第二の実施例について説明する。

【0039】本実施例は、検索条件として指定された文章あるいは文書から抽出した検索用の文字列を編集する際に、文書表示時に表示形式を変更する（以下、ハイライトと呼ぶ）文字列をユーザに選択させる方法である。本実施例により、ユーザが重要視している文字列のみが文書表示時にハイライトされるので、ユーザは文書中の重要箇所を瞬時に認識し、所望の文書か否かを高速に判別することが可能となる。

【0040】本実施例は基本的に第一の実施例（図1）と同様の構成をとるが、その中の文書検索制御プログラ

ム116および文書表示制御プログラム123が制御するプログラムの構成が異なる。文書検索制御プログラム116aが制御するプログラムの構成を図9に、文書表示制御プログラム123aが制御するプログラムの構成を図10に示す。文書検索制御プログラム116aはキーボード101からの文書検索指示により、システム制御プログラム111によって起動され、検索条件解析プログラム117、検索文字列編集プログラム900、表示形式変更文字列選択プログラム901、出現情報および重み取得プログラム119、類似度算出プログラム121および検索結果一覧表示プログラム122の制御を行う。文書表示制御プログラム123aはキーボード101からの文書表示指示により、システム制御プログラム111によって起動され、表示形式変更文字列取得プログラム1000、表示形式変更プログラム126および文書表示プログラム127の制御を行う。

【0041】以下、第一の実施例と異なる文書検索制御プログラム116aによる文書検索処理、文書表示制御プログラム123aによる文書表示処理の処理内容について説明する。

【0042】文書を検索する際には、キーボード101からの文書検索指示により、システム制御プログラム111が文書検索制御プログラム116aを起動し、文書検索制御プログラム116aが検索条件解析プログラム117、検索文字列編集プログラム900、表示形式変更文字列選択プログラム901、出現情報および重み取得プログラム119、類似度算出プログラム121および検索結果一覧表示プログラム122による一連の文書検索処理を制御する。文書検索制御プログラム116aは、まず文書検索の前準備として、磁気ディスク装置106に格納されている出現情報ファイル108と重みファイル109を主メモリ110に確保されたワークエリア128に読み込む。そして、文書検索処理では、図11のPAD図に示すように、まずステップ1100で、検索条件解析プログラム117を起動し、キーボード101から入力された検索条件を解析する。ここで、入力された検索条件には文章または文書が指定されていると、ステップ1101で判断された場合には、ステップ1102で検索文字列抽出プログラム118を用いて、検索条件として指定された文章または文書から、予め定められた文字列を抽出する。また、入力された検索条件には複数の単語の列が指定されていると、ステップ1101で判断された場合には、ステップ1103で検索条件から単語の列を抽出する。以上の処理内容は第一の実施例と同様である。次にステップ1104で、検索文字列編集プログラム900を起動し、検索条件から抽出した文字列を表示する。そして、表示した文字列に対して、ユーザに検索に用いる文字列の追加、削除を行わせる。さらにステップ1105で、表示形式変更文字列選択プログラム901を起動し、編集結果の文字列から文

書表示時に表示形式を変更する文字列をユーザに選択させ、その内容をワークエリア128に記憶する。次にステップ1106で、出現情報および重み取得プログラム119を起動し、編集結果の文字列の出現情報および重みを、ワークエリア128に読み込んだ出現情報ファイル108と重みファイル109から取得する。そしてステップ1107で、類似度算出プログラム121を起動し、ステップ1106で取得した文字列の出現情報と重みを用いて、予め定められた算出方法で検索条件とテキストデータ間の類似度を算出する。最後にステップ1108で、検索結果一覧表示プログラム122を起動し、ステップ1107で算出した類似度の降順にテキストデータを整列し、検索結果一覧として表示する。ステップ1106からステップ1108の処理内容は第一の実施例と同様である。以上で文書検索処理は終了する。

【0043】文書を表示する際には、キーボード101からの文書表示指示により、システム制御プログラム111が文書表示制御プログラム123aを起動し、文書表示制御プログラム123aが表示形式変更文字列取得プログラム1000、表示形式変更プログラム126および文書表示プログラム127による一連の文書表示処理を制御する。文書表示処理の処理内容を図12のPAD図に示す。文書表示制御プログラム123aは、図12に示すように、まずステップ1200で、表示形式変更文字列取得プログラム1000を起動し、文書検索制御プログラム116aによる文書検索処理で記憶した表示形式を変更する文字列をワークエリア128から取得する。次にステップ1201で、表示形式変更プログラム126を起動し、表示対象として指定された文書（以下、表示文書と呼ぶ）の中で、ステップ1200で取得した文字列が含まれる部分の表示形式を変更する。この表示形式の変更方法は従来技術2と同様である。最後にステップ1202で、文書表示プログラム127を起動し、ステップ1201で表示形式を変更した表示文書を表示して文書表示処理を終了する。

【0044】以上が本実施例における処理内容の概要である。

【0045】以下、図11に示した文書検索制御プログラム116aによる文書検索処理、図12に示した文書表示制御プログラム123aによる文書表示処理の処理内容について、具体例を用いて詳細に説明する。

【0046】文書検索制御プログラム116aによる文書検索処理の内容は図11に示す通りである。以下、具体的に説明する。まずステップ1100で、検索条件解析プログラム117を起動し、キーボード101から入力された検索条件を解析する。ここで、入力された検索条件には文章または文書が指定されていると、ステップ1101で判断された場合には、ステップ1102で検索文字列抽出プログラム118を用いて、検索条件として指定された文章または文書から、予め定められた文字

列を抽出する。また、入力された検索条件には複数の単語の列が指定されていると、ステップ1101で判断された場合には、ステップ1103で検索条件から単語の列を抽出する。以上の処理内容は第一の実施例と同様である。次にステップ1104で、検索文字列編集プログラム900を起動し、検索条件から抽出した文字列（以下、検索文字列と呼ぶ）を表示する。そして、表示した検索文字列に対して、ユーザに検索に用いる文字列の追加、削除を行わせる。さらにステップ1105で、表示形式変更文字列選択プログラム901を起動し、編集結果の文字列から文書表示時に表示形式を変更する文字列（以下、表示文字列と呼ぶ）をユーザに選択させ、その内容をワークエリア128に記憶する。図13にステップ1104およびステップ1105における検索文字列編集および表示文字列選択の画面例を示す。本図に示す画面には、検索文字列の一覧が表示される。ここで、ユーザは検索に用いる文字列を選択し、検索用と表示されたチェックボックスをオンにする。この結果、チェックボックスをオンに設定した文字列が検索に用いられ、それ以外の文字列は検索対象から排除される。また、表示された検索文字列以外の文字列を検索に用いる場合には、“文字列の追加”ボタンを押して文字列追加指示を入力することにより、ユーザが希望する文字列を追加する。表示文字列の選択は、本画面において、表示用と表示されたチェックボックスをオンにすることにより実現できる。ここでチェックボックスをオンに設定した文字列が、文書表示時に表示形式を変更して表示される。ここでいう表示形式とは、例えば色、サイズ、フォントやスタイル（太字や斜体）等である。そして、全ての設定が終了した時点で、ユーザは“検索実行”ボタンを押して、検索実行指示を入力する。その結果、表示用のチェックボックスがオンに設定された文字列をワークエリア128に記憶し、ステップ1106からステップ1108を実行する。ステップ1106では、出現情報および重み取得プログラム119を起動し、編集結果の文字列、すなわち図13において検索用のチェックボックスがオンに設定された文字列の出現情報および重みを、ワークエリア128に読み込んだ出現情報ファイル108と重みファイル109から取得する。そしてステップ1107で、類似度算出プログラム121を起動し、ステップ1106で取得した文字列の出現情報と重みを用いて、予め定められた算出方法で検索条件とテキストデータ間の類似度を算出する。最後にステップ1108で、検索結果一覧表示プログラム122を起動し、ステップ1107で算出した類似度の降順にテキストデータを整理し、検索結果一覧として表示する。ステップ1106からステップ1108の処理内容は第一の実施例と同様である。以上で文書検索処理は終了する。

【0047】文書表示制御プログラム123aによる文書表示処理の内容は図12に示す通りである。以下、具

体的に説明する。まずステップ1200で、表示形式変更文字列取得プログラム1000を起動し、文書検索制御プログラム116aによる文書検索処理で記憶した表示形式を変更する文字列をワークエリア128から取得する。図13に示す検索文字列編集および表示文字列選択画面の例では、ユーザは表示用として“サッカー”、“W杯”、“会場”および“ワールドカップ”のチェックボックスをオンに設定したので、ワークエリア128からこれらの文字列が取得される。次にステップ1201で、表示形式変更プログラム126を起動し、表示対象として指定された文書（以下、表示文書と呼ぶ）の中で、ステップ1200で取得した文字列が含まれる部分の表示形式を変更する。この表示形式の変更方法は従来技術2と同様である。最後にステップ1202で、文書表示プログラム127を起動し、ステップ1201で表示形式を変更した表示文書を表示する。図14に文書番号“123”、文書番号“003”の文書表示例を示す。本図に示す例では、ステップ1200で取得した文字列“サッカー”、“W杯”、“会場”および“ワールドカップ”が含まれる部分のサイズが大きく、斜体、強調が施されて表示されている。以上で、文書表示処理を終了する。

【0048】以上説明したように、本実施例では、検索条件として指定された文章あるいは文書から抽出した検索用の文字列を編集する際に、文書表示時に表示形式を変更する文字列をユーザに選択させ、選択された文字列が含まれる部分の表示形式のみを変更している。文書や長い文章を指定して検索を行い、その結果として得られた文書を表示する際に、検索に用いた文字列全ての表示形式を変更すると、図3に示すように変更箇所が多くなり、かえって重要な部分が見つけにくくなるという問題があるが、本実施例に示す方法を用いることにより、ユーザが重要視している文字列のみに限定し、表示形式を変更することができる。

【0049】このことにより、ユーザは文書中の重要箇所を瞬時に認識し、所望の文書か否かを高速に判別することが可能となる。その結果、検索結果文書を閲覧する際のユーザインターフェースを向上することができる。

【0050】本実施例では、ユーザが選択した文字列の表示形式を変更する場合の例について説明したが、表示形式を変えるだけでなく、ユーザが選択した文字列に関する情報、例えば出現頻度等を一覧として表示することも上記と同様の処理で実現することが可能である。

【0051】また本実施例では、ユーザに表示形式を変更する文字列を選択させる場合の例について説明したが、検索文字列を編集する際にユーザが追加した文字列の表示形式を変えることも上記と同様の処理で実現可能である。追加された文字列は、ユーザが重要視している文字列であるため、文書表示時にこの文字列の表示形式を変更することにより、文書中の重要箇所を瞬時に認識

10

20

30

40

50

することができる。

【0052】また、指定された検索条件が、以前の検索結果一覧から選択された検索結果文書である場合に、その文書を検索するのに用いた検索条件中の文字列の表示形式を変えることも上記と同様の処理で実現可能である。例えば、単語を指定して検索を行い、その結果得られた文書を用いてさらに検索を行う場合では、最初に指定した単語は重要な単語である。文書表示時にこの単語の表示形式を変更することにより、文書中の重要箇所を瞬時に認識することができる。

【0053】以下、本発明の第三の実施例について説明する。

【0054】本実施例は、文書表示時に、その文書の得点算出に大きく寄与した文字列の表示形式を変更する（以下、ハイライトと呼ぶ）方法である。本方式を用いることにより、表示文書の得点がどのような文字列によって算出されたものであるかをユーザは瞬時に認識できるので、所望の文書か否かを高速に判別することが可能となる。

【0055】本実施例は基本的に第一の実施例（図1）と同様の構成をとるが、その中の文書表示制御プログラム123が制御するプログラムの構成が異なる。文書表示制御プログラム123bが制御するプログラムの構成を図15に示す。文書表示制御プログラム123bはキーボード101からの文書表示指示により、システム制御プログラム111によって起動され、出現情報取得プログラム1500、文字列毎類似度算出プログラム1501、表示形式変更文字列抽出プログラム125、表示形式変更プログラム126および文書表示プログラム127の制御を行う。

【0056】以下、第一の実施例と異なる文書表示制御プログラム123bによる文書表示処理の処理内容について説明する。

【0057】文書を表示する際には、キーボード101からの文書表示指示により、システム制御プログラム111が文書表示制御プログラム123bを起動し、文書表示制御プログラム123bが出現情報取得プログラム1500、文字列毎類似度算出プログラム1501、表示形式変更文字列抽出プログラム125、表示形式変更プログラム126および文書表示プログラム127による一連の文書表示処理を制御する。文書表示処理の処理内容を図16のPAD図に示す。文書表示制御プログラム123bは、図16に示すように、まずステップ1600で、出現情報取得プログラム1500を起動し、文書検索制御プログラム116による文書検索処理でワークエリア128に記憶した文字列とその重みを取得する。そして、表示対象として指定された文書（以下、表示文書と呼ぶ）を走査して、取得した文字列の表示文書における出現情報を取得する。この出現情報は文書登録制御プログラム112による文書登録処理で抽出される

出現情報と同じものである。従来技術1の場合には、文書のベクトルを作成する際に必要となる文字列の出現回数を抽出する（表示文書1件に対する処理のみを行うので、出現情報として文書番号は不要）。次にステップ1601で、文字列毎類似度算出プログラム1501を起動し、ステップ1600で取得した出現情報と重みを用いて、表示文書の類似度への文字列毎の寄与率を算出する。そして、この寄与率の降順に文字列を整理する。表示文書の類似度は文書検索制御プログラム116による文書検索処理で算出される、検索条件の内容と類似する度合いを表した数値である。一般的には文字列毎に出現情報や重みを用いて所定の方法で算出した値の総和を類似度とするため、この文字列毎に算出した値の類似度に対する割合を寄与率とすることができる。次にステップ1602で、表示形式変更文字列抽出プログラム125を起動し、ステップ1601で整理した文字列の上位m個（mは予め定められた1以上の整数）を抽出する。このmの値はシステムで自動的に適切な値を設定しても良いし、ユーザに予め設定させておいても良い。また、文書表示毎に対話的にユーザに設定させ、適切な値を調整してもかまわない。さらに寄与率のしきい値を決め、しきい値以上の寄与率の文字列を抽出しても良い。次にステップ1603で、表示形式変更プログラム126を起動し、表示文書の中で、ステップ1602で抽出した文字列が含まれる部分の表示形式を変更する。この表示形式の変更方法は従来技術2と同様である。最後にステップ1604で、文書表示プログラム127を起動し、ステップ1603で表示形式を変更した表示文書を表示して文書表示処理を終了する。

【0058】以上が本実施例における処理内容の概要である。

【0059】以下、図16に示した文書表示制御プログラム123bによる文書表示処理の処理内容について、具体例を用いて詳細に説明する。まずステップ1600で、出現情報取得プログラム1500を起動し、文書検索制御プログラム116による文書検索処理でワークエリア128に記憶した文字列とその重みを取得する。そして、表示文書を走査して、取得した文字列の表示文書における出現情報を取得する。この出現情報は文書登録制御プログラム112による文書登録処理で抽出される出現情報と同じものである。従来技術1の場合には、文書のベクトルを作成する際に必要となる文字列の出現回数を抽出する（表示文書1件に対する処理のみを行うので、出現情報として文書番号は不要）。図17に本処理の処理例を示す。検索に用いられた文字列“W杯”、“サッカー”、“会場”、“来月”、“決定”、…とその重みをワークエリア128から取得し、これらの文字列で表示文書を走査することにより出現情報を取得する。本図に示す例では、出現情報として文書における出現回数をを用いている。次にステップ1601で、文字列

毎類似度算出プログラム1501を起動し、ステップ1600で取得した出現情報と重みを用いて、表示文書の類似度への文字列毎の寄与率を算出する。そして、この寄与率の降順に文字列を整列する。表示文書の類似度は文書検索制御プログラム116による文書検索処理において算出される、検索条件の内容と類似する度合いを表した数値である。一般的には文字列毎に出現情報や重みを用いて所定の方法で算出した値の総和を類似度とするため、この文字列毎に算出した値の類似度に対する割合を寄与率とすることができる。図17の例では、“W杯”、“サッカー”、“会場”、“来月”、“決定”、…の重みと表示文書における出現回数から、表示文書の類似度への文字列毎の寄与率を算出する。そしてこの寄与率の降順に文字列を整列することにより、“サッカー”、“W杯”、“会場”、“開催”、“決定”、…が得られる。次にステップ1602で、表示形式変更文字列抽出プログラム125を起動し、ステップ1601で整列した文字列の上位m個（mは予め定められた1以上の整数）を抽出する。このmの値はシステムで自動的に適切な値を設定しても良いし、ユーザに予め設定させておいても良い。また、文書表示毎に対話的にユーザに設定させ、適切な値を調整してもかまわない。さらに寄与率のしきい値を決め、しきい値以上の寄与率の文字列を抽出しても良い。図18に本処理の処理例を示す。本図に示す例では、上位3個、すなわちmの値を3としている。その結果、表示形式を変更する文字列として“サッカー”、“W杯”および“会場”が抽出される。そしてステップ1603で、表示形式変更プログラム126を起動し、表示文書の中で、ステップ1602で抽出した文字列が含まれる部分の表示形式を変更する。この表示形式の変更方法は従来技術2と同様である。最後にステップ1604で、文書表示プログラム127を起動し、ステップ1603で表示形式を変更した表示文書を表示する。図18に示す文書123の例では、表示形式を変更する文字列“サッカー”、“W杯”および“会場”が含まれる部分のサイズが大きく、斜体、強調が施されて表示されている。また本図には、文書003を表示する場合の例も示している。文書003が表示文書として指定された場合には、類似度への寄与率上位3個の文字列として“サッカー”、“準備”および“来月”が抽出され、これらの文字列が含まれる部分の表示形式を変更して文書が表示される。以上で、文書表示処理を終了する。

【0060】以上説明したように、本実施例では、文書表示時に、その文書の得点算出に大きく寄与した文字列を抽出し、抽出した文字列が含まれる部分の表示形式のみを変更している。文書や長い文章を指定して検索を行い、その結果として得られた文書を表示する際に、検索に用いた文字列全ての表示形式を変更すると、図3に示すように変更箇所が多くなり、かえって重要な部分が見

つけにくくなるという問題がある。しかし、本実施例に示す方法を用いることにより、表示文書の得点がどのような文字列によって算出されたものであるかをユーザは瞬時に認識できるので、所望の文書か否かを高速に判別することが可能となる。その結果、検索結果文書を閲覧する際のユーザインターフェースを向上することができる。

【0061】本実施例では、文書の得点算出に大きく寄与した文字列の表示形式を変更する場合の例について説明したが、表示形式を変えるだけではなく、得点算出に寄与した文字列の情報、例えば寄与率等を一覧として表示することも上記と同様の処理で実現可能である。所望の文書ではなかった文書の得点算出に寄与した文字列を検索用の文字列から削除することにより、この文書の得点を小さくし、所望の文書の得点を相対的に大きくすることにより、高速に所望の文書を入手することができる。

【0062】以上の実施例によれば、検索に用いられた文字列の数が大量となる場合でも、検索条件の内容と類似する可能性を示す指標に対して影響する文字列を選定し、それに関する情報のみを表示することができるため、ユーザは文書の重要箇所を瞬時に認識し、所望とする文書か否かを高速に判別することが可能となる。この結果、文書や長い文章を指定して検索を行い、その結果として得られた文書を表示する場合でも、所望の文書か否かの判別が容易な文書表示インターフェースを持つ文書検索システムを実現することができる。

【0063】

【発明の効果】本発明によれば、所望の文書か否かの判別が容易な文書検索方法およびシステムを実現することができる。

【図面の簡単な説明】

【図1】本発明の第一の実施例の構成を示す図

【図2】従来技術1の処理内容を示す図

【図3】従来技術2の問題点の説明図

【図4】本発明の文書表示処理の内容を示す図

【図5】文書登録制御プログラム112の処理内容を示すPAD図

【図6】文書検索制御プログラム116の処理内容を示すPAD図

【図7】文書表示制御プログラム123の処理内容を示すPAD図

【図8】出現情報ファイル108と重みファイル109の作成例を示す図

【図9】本発明の第二の実施例における文書検索制御プログラム116aの制御下のプログラムの構成を示す図

【図10】本発明の第二の実施例における文書表示制御プログラム123aの制御下のプログラムの構成を示す図

【図11】文書検索制御プログラム116aの処理内容

を示すPAD図

【図12】文書表示制御プログラム123aの処理内容を示すPAD図

【図13】検索文字列編集および表示文字列選択画面例を示す図

【図14】本発明の第二の実施例の文書表示処理の内容を示す図

【図15】本発明の第三の実施例における文書表示制御プログラム123bの制御下のプログラムの構成を示す図

【図16】文書表示制御プログラム123bの処理内容を示すPAD図

【図17】本発明の第三の実施例の出現情報取得処理、文字列毎類似度算出処理の内容を示す図

【図18】本発明の第三の実施例の表示形式変更文字列抽出処理、表示形式変更処理、文書表示処理の内容を示す図

【符号の説明】

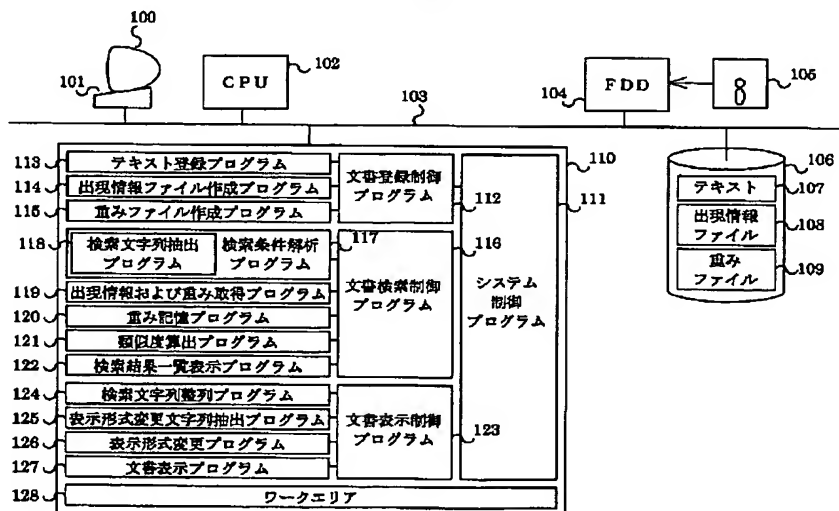
100・・・ディスプレイ、
101・・・キーボード、
102・・・CPU、
103・・・バス、
104・・・フロッピディスクドライバ、
105・・・フロッピディスク、
106・・・磁気ディスク装置、
107・・・テキスト、
108・・・出現情報ファイル、
109・・・重みファイル、

* 110・・・主記憶装置、
111・・・システム制御プログラム、
112・・・文書登録制御プログラム、
113・・・テキスト登録プログラム、
114・・・出現情報ファイル作成プログラム、
115・・・重みファイル作成プログラム、
116・・・文書検索制御プログラム、
116a・・・文書検索制御プログラム、
117・・・検索条件解析プログラム、
118・・・検索文字列抽出プログラム、
119・・・出現情報および重み取得プログラム、
120・・・重み記憶プログラム、
121・・・類似度算出プログラム、
122・・・検索結果一覧表示プログラム、
123・・・文書表示制御プログラム、
123a・・・文書表示制御プログラム、
123b・・・文書表示制御プログラム、
124・・・検索文字列整列プログラム、
125・・・表示形式変更文字列抽出プログラム、
126・・・表示形式変更プログラム、
127・・・文書表示プログラム、
128・・・ワークエリア、
900・・・検索文字列編集プログラム、
901・・・表示形式変更文字列選択プログラム、
1000・・・表示形式変更文字列取得プログラム、
1500・・・出現情報取得プログラム、
1501・・・文字列毎類似度算出プログラム

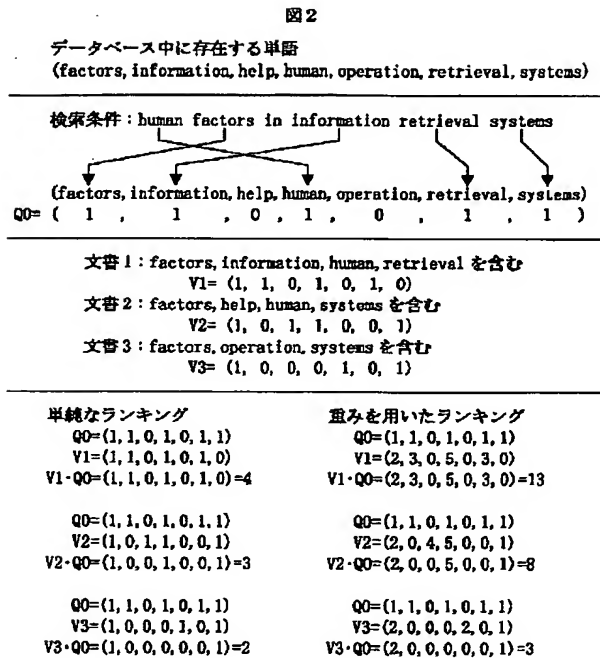
*

【図1】

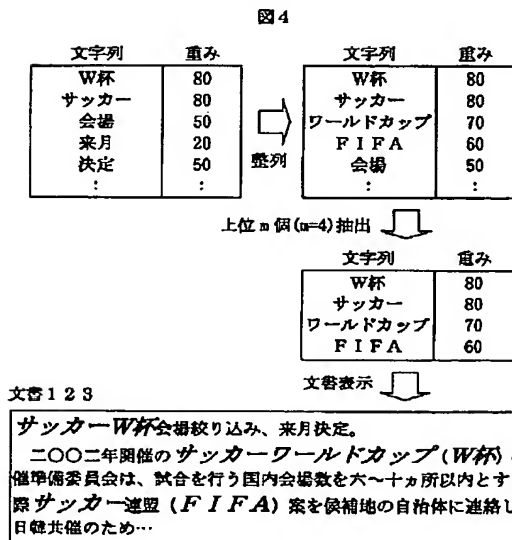
図1



【図2】



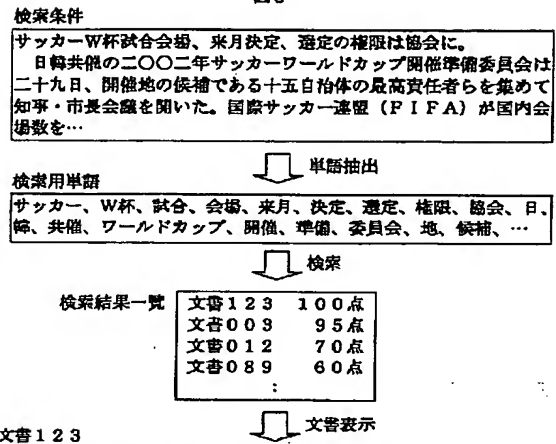
【図4】



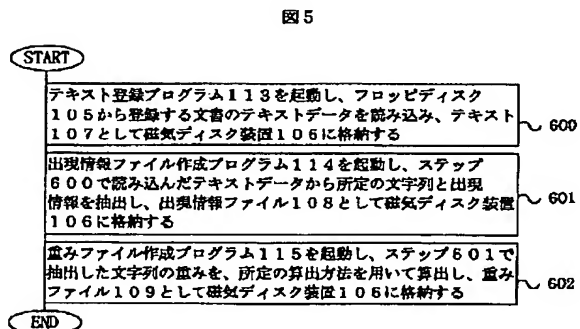
文書003

山形県、JFL加盟のサッカーチーム開設準備。
山形県サッカー委員会は、JFL加盟のサッカーチームの設立を決定し、準備室を来月開設する。同準備室を窓口として、サッカーチームの設立へ向けた出資者の募集や試合会場の選定などの準備をする。会場の候補として、...

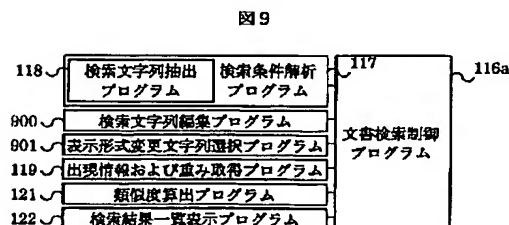
【図3】



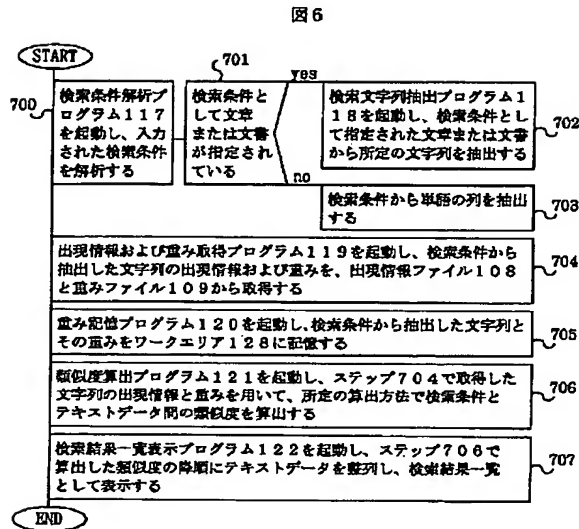
【図5】



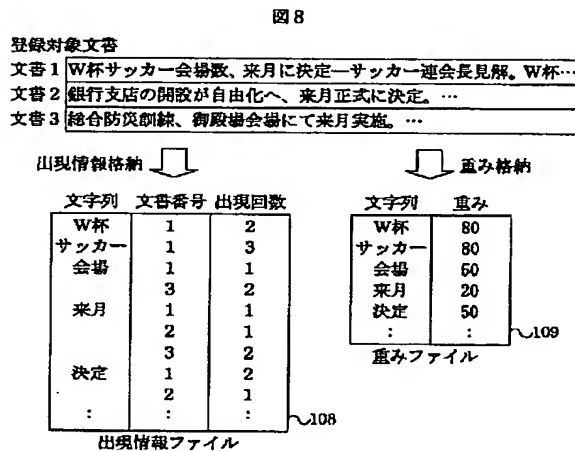
【図9】



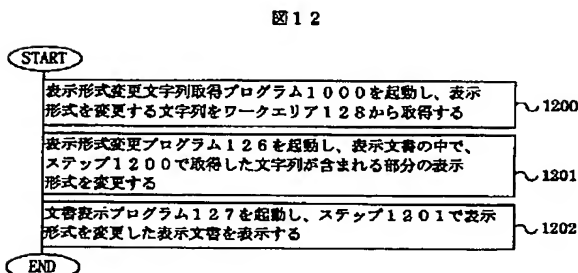
【図6】



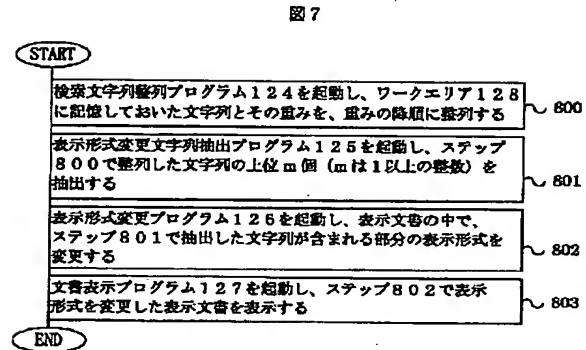
【図8】



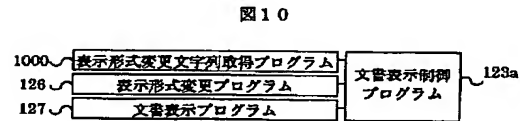
【図12】



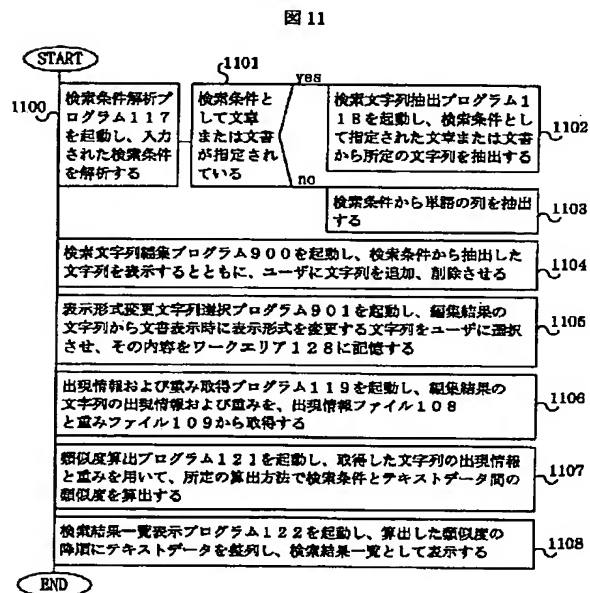
【図7】



【図10】

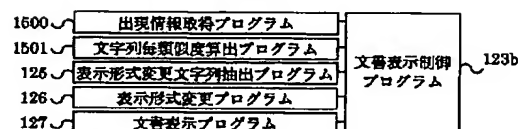


【図11】



【図15】

図16



【図13】

図 13

検索文字列編集および表示文字列選択画面

検索実行
文字列の追加

表示用 検索用

<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	サッカー
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	W杯
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	試合
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	会場
<input type="checkbox"/>	<input type="checkbox"/>	来月
<input type="checkbox"/>	<input checked="" type="checkbox"/>	決定
<input type="checkbox"/>	<input checked="" type="checkbox"/>	選定
<input type="checkbox"/>	<input type="checkbox"/>	権限
<input type="checkbox"/>	<input checked="" type="checkbox"/>	協会
<input type="checkbox"/>	<input type="checkbox"/>	日
<input type="checkbox"/>	<input type="checkbox"/>	輪
<input type="checkbox"/>	<input type="checkbox"/>	共催
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	ワールドカップ
		...

サッカー

W杯

試合

会場

来月

決定

選定

権限

協会

日

輪

共催

ワールドカップ

...

【図14】

図 14

取得文字列

サッカー

W杯

会場

ワールドカップ

文書表示

文書123

サッカーW杯会場絞り込み、来月決定。

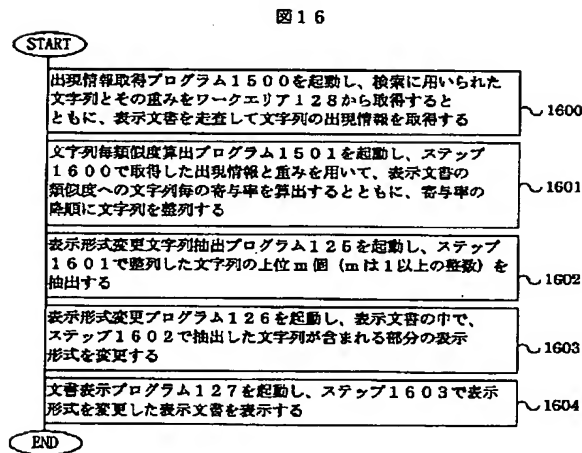
二〇〇二年開催のサッカーワールドカップ(W杯)の開催準備委員会は、試合を行う国内会場数を六〜十ヵ所以内とする国際サッカー連盟(FIFA)案を候補地の自治体に連絡した。日韓共催のため...

文書008

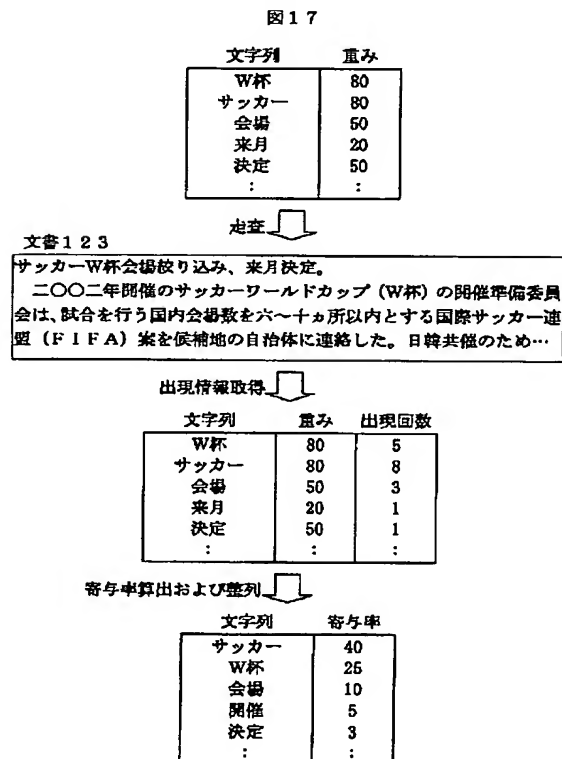
山形県、JFL加盟のサッカーチーム開設準備。

山形県サッカー委員会は、JFL加盟のサッカーチームの設立を決定し、準備室を来月開設する。同準備室を窓口として、サッカーチームの設立へ向けた出資者の募集や試合会場の選定などの準備をする。会場の候補として、...

【図16】

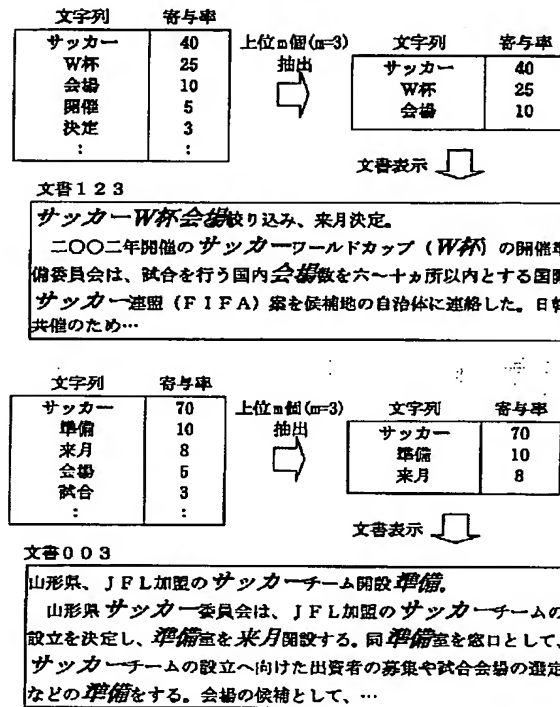


【図17】



【図18】

図18



フロントページの続き

(72)発明者 松林 忠孝
神奈川県川崎市幸区鹿島田890番地 株式
会社日立製作所ビジネスソリューション開
発本部内

(72)発明者 山口 明彦
神奈川県川崎市幸区鹿島田890番地 株式
会社日立製作所ビジネスソリューション開
発本部内

(72)発明者 稲場 靖彦
神奈川県川崎市幸区鹿島田890番地 株式
会社日立製作所ビジネスソリューション開
発本部内

(72)発明者 後地 陽介
神奈川県横浜市戸塚区戸塚町5030番地 株
式会社日立製作所ソフトウェア事業部内
F ターム(参考) 5B075 ND03 PQ02 PQ22 PQ36 PR04
PR06 QM08 UU06